Digital Preservation: An Introduction

12 July 2002
Delos Summer School, PISA

Dr Seamus Ross
Director, Humanities Computing & Information Management
Humanities Advanced Technology and Information Institute--HATII
University of Glasgow
&
Director, ERPANET
What Will I talk about

• Background to Preservation
• Obstacles
• Mechanisms for Preservation
• Case study of Electronic Records (e.g. email & digital documents)
• What is Authenticity
What’s Happening in Preservation

- Collaborative Projects—
  - From Pittsburgh, InterPARES, NEDLIB, to CEDARS/CAMiLEON, Presto

- National Initiatives led by Libraries & Archives
  - Such as ….. UK (NDAD), NEDLIB

- Gap between commercial activity and the knowledge in the public sector about these

- Legal Challenges to digital preservation (IPR, privacy) but FOI & legal requirements

- Recognition that our cultural memory is at risk and that it is composed of many types of digital objects (e.g. audio, VR)
Documents & Records = Institutional Assets

Maintaining the Accessibility & Reusability of Intellectual Capital

Storage Access Authenticity
Digital resources & problems

- Retroconversion of analogue materials (e.g. print, sound, images)
- Online resources: electronic journals, books, newsletters
- Databases, image & sound banks
- Multimedia productions

Fundamental Problem:
Specially created resources as opposed to purchased products delivered on portable media.
What kinds of resources

• discovery resources
  – such as catalogues, dictionaries, and bibliographies;

• digitised materials
  – such as audio, moving images, manuscripts, printed sources (e.g. online journals, books, newsletters)

• research resources and databases
  – such as sound and image banks, population censuses and surveys
Resource Creation Projects

• NINCH Review of Digital Imaging Practice shows insufficient attention to sustainability and to preservation issues
Difficulties Facing Creators and Users

- What information should be retained?
- Where should it be stored?
- What about the diversity of document types?
- How do I access it if I need it?
- How long should it be kept?
- What is its value?
- What are the costs and justifications?
- Does record creation equal retention?
Preservation questions

- What should be archived?
- What levels of documentation will be required?
- What selection criteria should be used?
- What standards should be used?
- Who pays? Who uses? Who selects?
- Medium, environment, context, integrity
Appraisal and Selection

- Administrative value
- Evidential value
  - (e.g. product liability)
- Informational value
- Reusability & Integration
- Technical viability
- Anticipated costs of preservation
- Usage restrictions

RM involvement in system design stage essential
Access, Intelligibility, and Maintainability

- Topology of data/information resources
- Hardware & software issues
- Migration and selection
- Storage strategies
- Migration and preservation infrastructure

Materials must be identified for preservation before they are created if activities, processes and systems are to support their preservation.
Recurring Value of Digital Objects

- Industry dependent
- Product liability
- Competitive advantage
- Recurring value through reuse
- Commercially valuable information a candidate for preservation
- Corporate memory
- Costs of re-creation vs storage
- Foundation for scholarly endeavour
Information Risks

- Uncontrolled growth in data and records
- Possibility of accidental record loss
- Security risks and information leakage
- Record duplication and authentication
- Unauthorised modification of records
- Loss of integrity and authenticity of digital resources
Objective of digital preservation

“retaining the ability to display, retrieve, manipulate, and use digital information in the face of constantly changing technology”
Making the Information & Knowledge Environment Work

- If the national and global network infrastructures are to provide a suitable business environment then systems must be in place to guarantee that the requirements for: integrity, authenticity, reliability and the archiving of digital information can be carried out easily and effectively.
Attentiveness to materials

- Access to digital materials depends on continued attention and maintenance
- Process requires close monitoring and must be controlled
- Interchangability of media, file formats,
- Peripheral devices
- Avoid propriety standards for encoding, software, & hardware
**Organisational Obstacles to Preservation**

- Tendency towards decentralisation & networked organisational structures
- Lack of collaboration between records managers, creators, and IT staff
- Need to link records management strategies with organisational objectives
- Lack of organizational commitment (social, economic, political)
- Failure to acknowledge the necessary [large] investment
- Failure to identify recognizable benefits
- Failure to link Preservation to corporate obj. s.
Key Preservation Issues

- **Medium**
  - storage media naturally decay
- **Technological (e.g. hardware/software)**
  - hardware and software obsolescence makes data/information inaccessible
- **Intellectual**
  - validation of integrity and authenticity
- **Contextual**
  - avoid loss of meaning with metadata
- **Legal Impediments**
Obsolescence & degradation

- Hardware (including access devices)
- Software
  - Operating Systems
  - Device drivers
  - Applications
- Media developments & degradation
- Contextual divergence
- Legal impediments
- Documentation & system divergence
- Distributed Networks
Obstacles to accessing surviving digital resources

- Loss of functionality of access devices (e.g. lack of drivers or interface functionality)
- Media degradation (e.g. temp & hum, disaster, manufacturer defects)
- Loss of manipulation capabilities (e.g. hardware, software, applications)
- Loss of presentation capabilities
- Weak links in creation chain (capture, manipulation, storage, dissemination)
Media type diversity

- Paper-based media
  - Punched Cards
  - Paper Tape
- Magnetic Media
- Optical Media
- Magneto-Optical
- Flash Cards
- etc.......

It is clear that the technology will become obsolete long before the data on the medium are lost.
Common types of magnetic media

- Floppy disks (e.g. 8”, 5.25”, 3.5”, 2” & density of storage)
- Removable disc packs (e.g. Bernoulli)
- Hard disks (1984 10 MB to 2001 73 GB)
- Magnetic tape
Magnetic tape

- Nine-track
- Half-inch data cartridges
- Digital linear tape (DLT)
- Quarter inch Cartridges (QIC)
- Digital Audio Tape (DAT)
Variety of Magnetic Media Formats

The most common archival degradation of tape media is related to changes in polymer binding system that leads to excessive oxide shedding and head contamination”

IBM STORAGE SYSTEMS
Seamus Ross--'An Introduction to Digital Preservation'

Pigment - Loss of magnetic signal
- Deterioration of particle
- Self-demagnetization

Binder - Polyester Polyurethane Hydrolysis
- "Sticky" tape phenomenon
- Debris/head clogs
- Lubricant Loss

Substrate - Dimensional Changes
- Mis-tracking

Binder
Top Coat
Substrate (PET)
Back Coat
Pigment
Few magnetic media coating

- Magnetic media coating formulations changing (licensed to coating companies)
  - binder
  - ferromagnetic particles (Fe$_2$O$_3$, BaFe, CrO$_2$, Co-Fe$_2$O$_3$)
  - lubricant
- Few organisations that actually coat
- Variation in the coating quality and process
- Measures of quality: adhesion, abrasivity, durability, chemical stability, error rates
What can go wrong with Magnetic Media?

- Hydrolysis
- Binder breakdown
- Particle breakdown
- Loss of lubricant
- Deformation
## Coercivity of Magnetic Media

<table>
<thead>
<tr>
<th>Media Category</th>
<th>Oersteds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floppy disk 5.25”, 360k</td>
<td>300 Oe</td>
</tr>
<tr>
<td>Floppy disk 5.25”, 1.2MB</td>
<td>675 Oe</td>
</tr>
<tr>
<td>Floppy disk 3.5”, 720k</td>
<td>300 Oe</td>
</tr>
<tr>
<td>Floppy disk 3.5” 1.44MB</td>
<td>700 Oe</td>
</tr>
<tr>
<td>PC, Mini, Hard disk</td>
<td>550 Oe</td>
</tr>
<tr>
<td>1980s hard disks</td>
<td>900 – 1400 Oe</td>
</tr>
<tr>
<td>1990s hard disks</td>
<td>1400 – 2200 Oe</td>
</tr>
<tr>
<td>Mainframe spool ½”</td>
<td>310 Oe</td>
</tr>
<tr>
<td>Cartridge</td>
<td>550; 650 Oe</td>
</tr>
<tr>
<td>TK50</td>
<td>1500 Oe</td>
</tr>
<tr>
<td>Reel ½” or 1”</td>
<td>310 Oe</td>
</tr>
<tr>
<td>Cassette</td>
<td>675; 750 Oe</td>
</tr>
<tr>
<td>8mm/4mm</td>
<td>1050; 1500 Oe</td>
</tr>
<tr>
<td>Cartridge 3840</td>
<td>300 Oe</td>
</tr>
<tr>
<td>¼ QIC Tape (DC600A)</td>
<td>550 Oe</td>
</tr>
<tr>
<td>Credit Card Strip</td>
<td>600 Oe</td>
</tr>
<tr>
<td>Library ticket</td>
<td>600 Oe</td>
</tr>
<tr>
<td>Cassette VHS</td>
<td>675; 700 Oe</td>
</tr>
</tbody>
</table>
Factors limiting media life & its contents

- Manufacturer quality (BASF)
- Handling (e.g. spooling & tension)
- Storage conditions (e.g. temp, hum)
- Format (helical vs. longitudinal)
- Use
- Condition of peripheral devices
- Disaster & environmental conditions (e.g. pollution)
- Market penetration
### Factors limiting tape life

<table>
<thead>
<tr>
<th>Head parameters</th>
<th>Tape parameters</th>
<th>System parameters</th>
<th>Circumstance parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>material</td>
<td>structure</td>
<td>tension</td>
<td>temperature</td>
</tr>
<tr>
<td>hardness</td>
<td>magnetic powders</td>
<td>relative velocity</td>
<td>humidity</td>
</tr>
<tr>
<td>direction of crystal</td>
<td>abrasive</td>
<td>head protrusion</td>
<td></td>
</tr>
<tr>
<td>grain size</td>
<td>lubricant</td>
<td>contact pressure</td>
<td></td>
</tr>
<tr>
<td>protective layer</td>
<td>surface roughness</td>
<td>running time</td>
<td></td>
</tr>
<tr>
<td>head structure</td>
<td>surface finishing</td>
<td>multiple passes</td>
<td></td>
</tr>
<tr>
<td>contact area</td>
<td>stiffness</td>
<td>internal temp</td>
<td></td>
</tr>
<tr>
<td></td>
<td>thickness</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Magnetic Tape
- I-D1
- Data D-2
- Data D-3
- 3480
- 3490/3490e
- DLT
- Data 8mm / Data VHS
- DDS / 4mm
- QIC / QIC-wide
- CD-ROM
- WORM
- CD-R
- MO

### Optical Disk
- Newspaper (high lignin)
- High Quality (low lignin)
- "permanent" (buffered)

### Paper
- "Permanent" Film

### Microfilm
- Medium-Term Film
- Archival Quality (Silver)

<table>
<thead>
<tr>
<th>Length of Storage: based on products available in 1995</th>
<th>Magnetic Tape</th>
<th>Optical Disk</th>
<th>Paper</th>
<th>Microfilm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 week</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 weeks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 month</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: John van Bogart, National Media Laboratory--http://www.nml.org

Seamus Ross--'An Introduction to Digital Preservation'
Magnetic discs drives
Platters, cylinders, & tracks

Disc spins at upwards of 7000 RPM.

Read-write heads.

Images © IBM
Density increase against decreases in costs

<table>
<thead>
<tr>
<th>Year</th>
<th>Media</th>
<th>bpi/bypi*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1963</td>
<td>reel</td>
<td>200</td>
</tr>
<tr>
<td>1965</td>
<td>reel</td>
<td>556</td>
</tr>
<tr>
<td>1969</td>
<td>reel</td>
<td>800</td>
</tr>
<tr>
<td>1971</td>
<td>reel</td>
<td>1600</td>
</tr>
<tr>
<td>1981</td>
<td>reel</td>
<td>6250</td>
</tr>
<tr>
<td>1985</td>
<td>cartridge</td>
<td>37871*</td>
</tr>
<tr>
<td>1991</td>
<td>cartridge</td>
<td>77000*</td>
</tr>
</tbody>
</table>

Disc Storage
1956 $1 million for 5 megabytes
1998 $500 for 5 gigabytes
2002 $500 for 73 gigabytes

Although high-end tape technologies such as 18mm have very high initial costs, the cost per megabyte can be much lower than technologies with lower entry costs, such as DAT and 8mm.

© Freeman Associates
Optical Discs

• Read-only
  – Video Discs (1978 12” & 8” sizes)
  – Compact Disk (CD)
  – Digital Video Versatile Disk (DVD)

• Read/Write
  – WORM - Write Once Read Many Times
  – CD-Rs
Optical R&W Media

- Ablative -  
  - tellurium coated disc which pits when hit with a laser (1 = pit)

- Dye-based -  
  - uses an organic dye which is pitted a laser

- Phase-change  
  - laser heating of the changes the properties of the media

Lasers record information by altering the light reflectance characteristics of the medium.
Optical & tape technologies

Optical Media
- Universality of format
- Reasonable storage densities
- Rapid file location, but slow transfer rates
- Potential durability of media
- Greater hardware reliability
- Massive market penetration = high standardisation

Tapes
- Massive storage capacities
- Exceptionally fast transfer rates
- Excellent error correction facilities
- Relatively cheap storage medium
- Diversity of standards, variable market penetration, low hardware
- Active storage management
  - Tape pack/wind quality
  - Tension and rewinding (constant torque)
CD-ROM structure

- Protective coat
- Reflective metal (Al)
- "Data" layer
- Substrate
- Laser path

Source: National Media Laboratory--http://www.nml.org
What can go wrong with CDs?

Environmental Impacts
- corrosive gases
- temperature
- humidity

Handling Impacts
- shocks
- abrasions
- scratches

Mechanical Impacts
- degradation of hardware

Exposure to UV Light

Is CD-R an Archival Medium?
All CDs & access devices not created equal...

- Variation in materials used in the raw media
- Variation in process of manufacturing
  - etching,
  - pressing,
  - CD-R recording
- Drives constructed from components sourced from different manufacturers
- Variability in device MTF
- Viability of repairing or replacing mechanical components
ANA for CDs & DVDs

Avoid
- damage to the upper and lower surfaces and edges of the disc
- scratching and contact with surfaces that might result in grease deposits (e.g. human hands)
- exposing discs to direct sunlight

Never
- attach or fix anything to the surface of the CDs
- write on any part of the disk other than the plastic area of the spindle

Always
- store media in jewel case or protective sleeve when not in use
- use low lint acid-free archival quality sleeves, if using sleeves
- wear gloves when handling the master CDs
**ANA for DLT**

- **Avoid**
  - placing the tapes near magnetic fields
  - moving the tapes about
- **Never**
  - stack the tapes horizontally
  - put adhesive labels on the top, side or bottom of cartridge
  - touch the surface of the tape
  - put a tape that has been dropped in a drive with first visually inspecting it to make certain that the tape has not been dislodge or moved
- **Always**
  - keep tape in its protective case when not in use
  - move tapes in their cases
  - store the tapes in appropriate environmental conditions
  - store the tapes vertically
Acting to avoid dangers

- Frequent access results in excessive wear-and-tear
- Transportation
- Cleanliness (finger print, smoke particles, dust)
- Stray magnetism
- Poor storage (e.g. environmental)
- Poor quality read-write devices
- Loss of documentation
Media recording issues

- High quality devices only
- Cleanliness of equipment (e.g. read-write heads)
- Mechanical working order
- Electrical alignment
- Beware of reader obsolescence
Documentation (media)

- Media type
- When media written
- Who by
- Format, density,
- Compression type (if any--see below)
- Device used
- How verified
- Where stored
Reading & Writing Media

- High quality devices only
- Cleanliness of equipment (e.g. read-write heads)
- Mechanical & electrical condition
- Obsolescence of access device
Problem Areas

- Writing & Reading digital media
- Encoding standards (e.g. NRZ)
- Decoding algorithms
- Compression (e.g. LZW)
- Logical vs physical storage
- Diversity of data type formats (e.g. word-processing, images, sound)
- Diversity of file formats (e.g. TIFF, AUI)
- Documentation
Encoding & decoding

- Algorithms which manage how the bits (010110101) are written to the media
  - Return to Zero
  - Non-Return to Zero (NRZ)
  - Alternative Mark Inversion
  - High Density Bipolar 3

- Algorithms which then decode to restore meaning to the bit stream.

NRZ: 1 bit change polarity and 0 bit no change
Compression

- Hardware compression (Avoid it)
- Lempel-Ziv-Welch (LZW)
- Audio interchange format (AIFC)
- Image compression formats (jpg)
Logical vs physical storage

- File allocation
  - contiguous addresses
  - linked list of sectors across surface of disk
  - indexed allocation

- Allocation method stores other data as well (error detection & correction info)

On magnetic disks files are not always written contiguously often they are fragmented across sectors.
File formats

- Text
  - raw text
- Proprietary
  - word processed
  - spreadsheets
- Databases
- Images (e.g., xbm, jpg, etc)
- Moving image
- Sound (e.g., aui)
Role of software

Table 1. Example of Character Sequences for Bolding Text

<table>
<thead>
<tr>
<th>Software Application</th>
<th>Begin Bold (in Hexadecimal)</th>
<th>End Bold (in Hexadecimal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wordstar® 6.0</td>
<td>02</td>
<td>02</td>
</tr>
<tr>
<td>Ami Pro® 3.1</td>
<td>3C 2B 21 3E</td>
<td>3C 2D 21 3E</td>
</tr>
<tr>
<td>HyperText Markup Language</td>
<td>3C 42 3E</td>
<td>3C 2F 42 3E</td>
</tr>
</tbody>
</table>

From: Cpt S Robertson’s study

Technological options

- Technology preservation
- Data migration
- Emulation of hardware and software environment
Preservation Options

- Hardware and software preservation
  - technically complex and expensive
- Software & Hardware emulation
  - practical (?)
- Data migration
  - can lead to data and information loss
  - can lead to loss of functionality
- Virtual Machines
- Binary Retargetable Code
  - Transmogrify Adaptable Preservation (TAP)
The Stasii Tapes

- Unknown format
- Limited documentation
- Poor storage conditions
- Availability of code books
Recovering F-16 Recorders
The Challenger Disaster

Images © NASA & IBM

Seamus Ross--‘An Introduction to Digital Preservation’
Hardware replication: Eniac

- Eniac-on-a-Chip

Replacing a bad tube meant checking among ENIAC's 19,000 possibilities.

Images © University of Pennsylvania

Seamus Ross--'An Introduction to Digital Preservation'
Edsac & its simulation

Images © Martin Campbell-Kelly
University of Warwick
Ferranti Pegasus Computer

Images © Science Museum, London
Pegasus computer & its PC simulation

Images © Science Museum, London
Migration strategies

- Sequence of tasks undertaken periodically
- Change media
- Converting format or encapsulating
- Incorporation standards
- Time & labour dependencies
- Costs vs. value
- Influenced by processes, systems, & best practice

El Archivo General de Indias
http://www.mcu.es/lab/archivos/index.html
Recovering Lost Information

- Data Recovery
  - rescue of material from damaged media
  - recovery of unknown materials
  - maintain a range of hardware & software to aid recovery
  - emulation of hardware
  - emulation of software

Data Loss Summary:
- Natural Disasters: 3%
- Software Corruption: 15%
- System Malfunction: 46%
- Human Error: 36%

Source: Ontrack 1996
Promising Activities --
Doom and Gloom no more

- Magnetic Force Microscopy (MFM) to recover data from magnetic media
- Hardware and software emulation
- Binary Retargetable code
Magnetic Force Microscopy

- Developed about 10 yrs ago
- Atomic Force Microscope (AFM)
Magnetic Force Microscopy & Crashed discs

By increasing the sensitivity of the imaging it is possible to see magnetic tracks in the damaged area.
Digital Archaeology

- It is very difficult to lose your data completely, although it may be very expensive to recover it. So act to avoid compromising your valuable resources, but don’t panic when all goes wrong because everything need not be lost.
Information resources: databases & multimedia

- Well-structured information resources generally composed of limited range of categories of data
- Increasing tendency towards complex information resources composed of a diversity of data types (e.g. images, sound), heavily software and hardware dependent
The Evolving Document

♦ transaction databases
♦ static documents which are composed of such elements as text, tables, and images;
♦ multimedia or data-rich documents such as the kinds of documents that we encounter in the networked environment using such technologies as the world-wide web or using www-based corporate intranets (including databases); and,
♦ dynamic documents which are dependent upon data that might have variable instantiations.
Why Retain E-records

- active business use (e.g. decision making support)
- later accountability and research
- efficiency of storage and access
- hardcopies do not fully represent record
- primacy of record

‘Records are created in anticipation of future disputes.’ Robert Williams, (Imagining World, 3.10)
So what is a record

- ‘information produced or received in the initiation, conduct or completion of an institutional or individual activity’
- ‘comprises content, context, and structure to provide’
- it is EVIDENCE
Digital Object viability

- Accuracy - demonstrate system functions as specified and produces immutable digital objects
- Authenticity - that the record is what it purports to be
- Completeness
  - content
  - context = circumstances of creation
  - structure = logical & physical attributes
Qualities of Electronic Resources/Records

- Authentic
- Comprehensive
- Accurate
- Reliable
- Unique
- Version-controlled
Authenticity

- Why Authenticity & Integrity
  - Origin
  - Completeness
  - Internal integrity
  - Contextual precision

- When is Authenticity Important
  - Information discovery - relevance
  - Information retrieval
  - Information use

Proof that it is unaltered from the original
Resource Viability

- Accuracy - demonstrate system functions as specified and produces immutable records
- Authenticity - it is what it purports to be
- Completeness
  - content
  - context = circumstances of creation
  - structure = logical & physical attributes
- Unique
- Version control for documents
- Intelligible
- can be processed
- Retrievable
Methods of ensuring Authenticity

- **Public methods**
  - copyright deposit
  - registration
  - metadata

- **Secure methods**
  - digital watermarking
  - stegonography
  - digital signatures
  - encryption
  - embedded active agents
Digital Object Verification Strategies

- Audit Trails
- Encryption strategies
- Digital Signatures
- ‘Snapshots of transient compound documents’
- Evaluation of data quality = accuracy & timeliness. Audit trails & system evaluation
- Digital watermarking
- Stegonography
- Wrappers
- Embedded active agents
Weak points in e-records

- Creation
- Processing
- Storage
- Retrieval/Destruction

Fundamental issues are:
security & retention strategies.
Control the process

- Access
  - who can do what
  - how is use tracked
  - system security
- Evaluation of accuracy
- Alterations (track them)
Audit trails

- Who used the system
- When did they use it or instigate the transaction
- What did they do when they used the system
- What was the result of the transaction

Audit trails should demonstrate accuracy and reliability of transaction and its record.
Access and usage

- Secure at all stages
  - capture
  - manipulation
  - storage
  - dissemination
- Functional (known processes)
- Controlled and monitored
- Performance and perceived benefits
Preparation strategies

- Design systems for long term accessibility (e.g. work with developers)
- Control, monitor, document, and audit migration
- Avoid proprietary systems (e.g. hardware, software, applications, standards)
- Avoid emerging technologies
Documentation & Metadata

- Information identifying the resource
- Terms of access
- Guidelines to open and read
- Details of how, when and why the record
- Clues to its authenticity and verification
- Evidence of its use
- Assistance with the meaning of the record
- Essentially: structure, context, [content], use history
System Documentation: Metadata Elements

- Logical and physical models of the system
- Information flow models
- Data flow diagrams
- Entity-relationship charts
- Process model descriptions
- Data dictionaries
- Information Resource Directory Systems
Metadata Studies (2 Examples)

- Pittsburgh Project
  - handle layer
  - terms and conditions layer
  - structural layer
  - contextual layer
  - content layer
  - use history layer

- Astra AB
  - identifying metadata
  - interpretation metadata
  - context defining metadata
  - transaction metadata
    - conducted context
    - performed transactions
  - event-log
Preservation Metadata

• Comprehensive metadata framework applicable to the digital preservation activity
  – RLG/OCLC Working Group on Preservation Metadata
  – New Zealand National Library Metadata Framework
**Metadata Consensus Gaps**

- Testing of model across organisational types
- Representation of business processes (including information flow)
- Layer of documentation to cover system documentation
- Incorporation of metadata guidelines into software
Role of standards

- Crucial for information exchange
- Critical for information migration
- Not a solution to preservation problem alone - only high-level solution to some technical obstacles
- 30 years experience demonstrates that standards have a lifecycle of their own
Info is Power

- Electronic data crucial in litigation
- Key issues:
  - the cost of response
  - integrity and authenticity of source
  - access
  - security
- Email a case example
What About Email

- Judges, jurors and others like email because
  - they believe it reflects the true feelings of the author; its informal character is its greatest strength in their view
- Expense of Access to it without clear planning
Email and recordness

• What about email
  – access
  – obsolescence
  – laws and regulations (DP, FOI, CP)
  – reliability
  – confidentiality

Email ‘unlike phone messages’ it is inherently archival and needs to be securely managed.
?? Print Out it Out ??

- Will Print outs of documents serve as evidence?
  - established key principal of records
- Public Citizen v Carlin
  - GRS20
Email - Sample Cases

- Armstrong vs Office of the President
- Public Citizen vs Carlin
- Long, Burnham, & Public Citizen vs Janet Reno and Executive Office of US Attorney’s Office
- NASA
- ATI v Sprint
  - ‘purposeful destruction of evidence because backup tapes not properly assessed’
- Shaw v Hughes
  - tape rotation led to the loss of evidence

(1) Proper management of E-recs essential.
(2) Secure methods of access must be in place.
NASA’s View

• Because all e-mail can be the target of a number of public and legal disclosure instruments, and as the government’s definition of ‘records’ is difficult to interpret and this policy is difficult to enforce, the agency has further stipulated that all email files (central store only) that are older than 60 days must be erased automatically.
GSLIS - Austin, Texas

• Surveyed industry specialists who agreed that:
  – organisations and individuals usually fail to consider the ‘evidential nature of email and rarer still [were] organisations which had an email policy.
  – many organisations were considering developing policy but not getting far with it
Not Uncommon Practice

- ‘Backup tapes containing electronic records from CMS in Western District of Kentucky ‘retained two weeks and overwritten’.
- Destruction of unique evidence was creating ‘irreparable and continuing harm’
Common thread to all cases

- Organisations must have a records/digital object schedule which provide rules for
  - retention
  - disposition of records
- Lack of records/digital object management programme
- lack of skills and resources not acceptable

‘identify and preserve necessary business docs while keeping the numbers of these to a minimum’.
Ensuring Recordness-ARMA

- Email messages handled as records
  - retention
  - responsibility
  - storage
- Email is not a single record series and has no single retention period
- use and content must be understood
- end-users must manage email
Ensuring Recordness-ARMA

- email exposes weaknesses in record keeping systems
  - what docs need to be filed
  - what drafts should be retained
  - how should access be provided
Document Types

- Annual Reports
- Balance Sheets
- Bibliographic or Catalog Information
- Blueprints or Architectural drawings
- Books
- Charts
- Contracts Software
- Databases
- Drawings Sketches
- E-Mail
- Graphics
- Handwritten notes
- Invitations
- Invoices
- Journals
- Labels & Indexes
- Legal documents
- Letters
- Magazines
- Manuscripts
- Maps
- Newspaper clippings & ads
- Payroll
- Phonorecords
- Photographs
- Presentations
- Program Components
- Proposals
- Receipts
- Research Notes
- Sheet Music
- Spread Sheets
- Standard A4 Posters
- Surveys
- Teleconference recordings
- Tests
- Tickets
- Transcripts
- VHS tapes

Seamus Ross--‘An Introduction to Digital Preservation’
Document Management

- Electronic Document Management (EDM) or Enterprise Document Management
- Document lifecycle
- Paper
- Imaging/scanning
- Optical Character Recognition (OCRing)
- Intranets
- Preservation, Integrity, Access
EDM vs RM

• Must be able to define whether a document is a record or not
• Formal retention and disposition scheduling is necessary
• It is necessary to track material outside the system
• Identify locations of records
<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Document Management</th>
<th>Records Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary application</td>
<td>Manage information in documents, make it easy to find</td>
<td>Control corporate assets, ensure compliance</td>
</tr>
<tr>
<td>focus</td>
<td>and access</td>
<td></td>
</tr>
<tr>
<td>What application is</td>
<td>Manage information in documents, make it easy to find</td>
<td>Control corporate assets, ensure compliance</td>
</tr>
<tr>
<td>designed to do</td>
<td>and access</td>
<td></td>
</tr>
<tr>
<td>Role of a 'document'</td>
<td>Information container</td>
<td>Evidence</td>
</tr>
<tr>
<td>Value basis</td>
<td>Re-usability, reference</td>
<td>Statutory, regulatory, operational, historic</td>
</tr>
<tr>
<td>Duration of usage</td>
<td>Instantaneous-used during creation, revision, or</td>
<td>Used during the entire life cycle</td>
</tr>
<tr>
<td></td>
<td>searching</td>
<td></td>
</tr>
<tr>
<td>General attitude</td>
<td>All information is created equal; keep everything</td>
<td>Information can be our undoing; destroy it as soon as</td>
</tr>
<tr>
<td></td>
<td></td>
<td>permitted</td>
</tr>
</tbody>
</table>
Digital Object & Document Software

- creates records which conform to the Duranti diplomatic guidelines
  - (eg authentic, comprehensive, accurate, reliable, unique, and version controlled)
- automated creation of record metadata
- supports automated attachment of use history
- produces application independent records
- copes with dynamic records
What should you do!

- Involve records/archivists in EDM development
- Incorporate training programmes in all initiatives
- Seek standard, portable solutions
- Define retention strategies during EDM development and automate them (5%)
- Plan for the future
- Read widely
Data Warehousing

- Subject-oriented
  - (e.g. customer, product, transaction, account)
- Integrated
  - introduces consistency across diversity
- Nonvolatile
  - warehoused data is never updated
- Time variant & time stamped
  - Opsys = 60-90 days vs 5 to 10 years
- Granularity of data
Data Warehousing vs Archiving

- Not a continuum for operational system to data warehouse to archives
- Archives hold data that is independently usable because it has data wrappers
- Reuse expectations not set for archives
- Frequency of usage often low
- Expected life = long term
- Retains original order and structure
- Maintains business process relationship
Archive characteristics

- Integrated into business objectives
- Holds reusable data
- Metadata secured data
- Has retention and destruction policies
- Has data migration strategies
- Centrally managed, but distributed
- Suitable access interfaces
- Maintains contextual currency
Record Retention Strategies - Canadian Models

- User driven
- Automated work processes
- Keep records electronically
- Automate record creation
- Desktop record access
- Disposition is automatic
- Training
**OAIS Model**

- **OAIS** = Open Archival Information Systems
- Key players in development
  - National Space Science Data Centre
  - Consultative Committee for Space Data Systems
- Premises Underlying OAIS
  - Data are irreplaceable (esp observation)
  - Data and associated metadata must be moved across technologies
  - Representations and formats will change
  - Lack of consensus on adequate metadata standards
Key OAIS Objectives

- Objective
  - recognised no framework for developing digital archive standards
  - need for a reference model
  - recognise the hybrid nature of archives
  - collaborate with archival community
  - focus on data resulting from space missions
  - near-term and indefinite storage of digital data
  - independent of implementation model
  - address full range of archival processes
Process of Development

- Examine other models
- Define Data Archiving
- Define functional areas (FAs) including ingest, storage, access, and preservation
- Define interfaces between FAs
- Define a set of data classes
- Formal representation methods
OAIS Overview

- Manages ingest of Information Packages from creators
- Defines the communities needing the Information
- Reflects needs of identified user community
- Enables preservation in an understandable way
- Uses documented policies and procedures
Advantages of OAIS

- Provides a model where one was lacking
- Facilitates procurement of systems
- Enables interoperability between OAIS complaint systems
- Supports the migration task
- Lays out a minimum set of responsibilities
Briefly what does it do?

- Information Definition
  - express as a data type
  - can be exchanged
- ‘Data interpreted using its Representation Information yields information’
- Preservation depends upon understanding the data object and associated representation information (DO + RI = IO)
- Defines and Information Package to contain: Content Information and Preservation Description Information (PDI)
What does it support?

- Archival Information Package
- Information Package Variants
  - Submission Information Package
  - Archival Information Package
  - Dissemination Information Package
- Relates activities of Producer, manager and consumer
- Supports Functional Entities
OAIS Functional Entities

- Ingest
- Archival Storage
- Data Management
- Administration
- Preservation Planning
- Access

(draw sketch to explain)
OAIS MODEL
Who is working with OAIS

- Archive & Library Community
  - Koninklijke Bibliotheek (KB) through NEDLIB—design and architecture of Deposit System for Electronic publications
  - CEDARS
  - NARA and the San Diego SuperComputer Center
  - National Space Science Data Center
  - Pharmaceutical & Aerospace Industries
  - French Space Agency for its plasma physics archive
InterPARES Preservation Model

From InterPARES Preservation Workgroup

Seamus Ross--'An Introduction to Digital Preservation'
InterPARES Preservation Model

From InterPARES Preservation Workgroup

Seamus Ross--'An Introduction to Digital Preservation'
InterPARES Preservation Model

From InterPARES Preservation Workgroup

Seamus Ross--"An Introduction to Digital Preservation"
InterPARES Preservation Model
InterPARES Preservation Model

From InterPARES Preservation Workgroup

Seamus Ross--'An Introduction to Digital Preservation'
InterPARES Preservation Model

From InterPARES Preservation Workgroup
InterPARES Preservation Model

From InterPARES Preservation Workgroup
InterPARES Preservation Model

From InterPARES Preservation Workgroup
Trusted Repositories

• What is one?
• RLG/OCLC Proposal
  – need a programme for certifying trusted repositories
  – checklist of concept and key elements needed
• Depends on definable, certified and auditable practices
• What would certification guarantee and how would it be revoked and with what implications
Aspects need certification

- People
  - through developing competencies
- data
  - Quality management, policy, validation
- processes
  - OAIS model, IPR, FOI, organisational practices
- managing organisations
  - audit of approaches organisations take to data management
Certification

- Statement of attributes to be measured
- Policies and Assumptions (e.g. practices, environment and security)
- Procedures against standards
- Relationship with depositors
- What processes are in place to manage fidelity checks for ingest
- What metadata processes are in place
- What user needs evaluation work is carried out
Acquisition of Experience

- Develop test experimental frameworks
- Experiment with ingesting, managing, and providing access digital assets
  - Netherlands: Digital Repository Project
  - US: NARA-San Diego Super Computer Center Project
- Do something concrete -- gain experience
- Ensure parameters of the research are well-documented so that they can be duplicated
- Aim for ‘recipe-like’ descriptions of processes
Cost Modelling

- Lack of credible cost models
- Little information about actual costs
- Few clear statements of the cost elements themselves and none that are comparable
Organisations Need Help

- Off-the-shelf policy statements
- Business cases & strategies
- Digestible guidance on technologies and their preservation implications
- Improved models (reference, costs, standards, functional requirements)
- Simple Guidelines on digital survival
- Access to Metadata Repositories
- Guidance on creating data repositories (see RLG/OCLC "Attributes of a Trusted Digital Repository"
- IPR support and guidance
Self-Aware Objects

- Digital objects that know what they are
- Digital objects that can observe the state of other objects (e.g. observe decline in numbers of similar classes of objects)
- Digital objects that know where they are
- Digital objects that know where their metadata are
- Digital objects that can notify their originator/manager if they need to be protected, migrated, secured
The Role of Audit

- System Audit
- Process Audit
- Archive Audit
- Access Audit
Trend from IT

- Information Technology has evolved to become Information, communications and technology
- Focus is now on information content rather than on technology
- Document management integral to any information strategy
- Records and resource management strategy integral to any information strategy
Problems with ICT Support

- Backup & Archiving
- Technology shift
- What does delete mean?
- How can you identify, locate, retrieve and preserve
Information Needs
- Access to information beyond public sector
- Dissemination (e.g. training workshops)

New Services
- An International Software Archive/Repository or a Software Museum
- Peripheral Devices Museum
- Public Sector Data Recovery Services

Research into....
- Access mechanisms and Evaluation methodologies
- Typology of digital information (We don’t know enough about digital assets)
- Emulation, Migration, Binary Retargetable Code
- Cryptography and magnetic force microscopy
Avoiding Digital Rainforests

- Acknowledge that information has value
- Define retention guidelines
- Automate the selection and archiving of records
- Establish metadata standards
- Develop record appraisal guidelines
- Improve communication
- Implement migration strategies
- People & training make solutions work
- Keep the bit stream.
Thinking Toward the Future

- Design information for long term accessibility (e.g. work with creators)
- Migration must be controlled, monitored, documented, and audited
- Avoid proprietary systems and emerging technologies
- Better software
- Intelligent record selection & appraisal tools
- Mechanisms for maintaining links between business process and records created/used by them
- Case studies of the cost-benefits analysis for data loss vs preservation